LANGUAGE-MODEL INVESTIGATIONS RELATED TO BROADCAST NEWS

Dietrich Klakow, Xavier Aubert, Peter Beyerlein, Reinhold Haeb-Umbach,
Meinhard Ullrich, Andreas Wendemuth and Patricia Wilcox

Philips Research Laboratories
Weisshausstr.2, D-52066 Aachen, Germany, klakow@pfa.research.philips.com

ABSTRACT

In this paper we present some experiments that have been performed while developing language models for the PHILIPS Broadcast News system. Three main issues will be discussed: construction of phrases, adaptation of remote corpora to this task, and the combination of the different models. Also, perplexities on the 1997 evaluation data are reported.

1. Introduction

Two main lines have been pursued to improve the system: construction of phrases and adaptation.

The first topic is the combination of words to phrases. It is questionable whether words are really the best basic units for the estimation of stochastic language models - grouping frequent word sequences to phrases can improve language models. This issue has already been raised elsewhere [1, 2]. Tests done on Wallstreet Journal (WSJ) and Broadcast News (BN) revealed that bigram perplexity (PP) can be reduced by up to 29%. In tests on WSJ we also observed reductions in word error rate (WER) of about 10%, which is partly due to language model (LM) and partly due to acoustic effects. From the LM point of view, phrases may be considered as a variant of varigrams [3].

The second topic is adaptation, based on the reliable estimation of unigram distributions. This method is related to work done at IBM [4, 5] and is described in full detail in [6]. In [7] more refined information about the specific domain is used. We present a method, which uses only the first iteration of the generalized iterative scaling algorithm [8] to combine the local unigram distribution and the background M-gram model. For planned speech the effect is small, however, for spontaneous speech (F1 and F2) this yields a perplexity reduction of more than 30% for the North American Business News based model. This technique was used to adapt and combine four different corpora: Broadcast News, British National Corpus (BNC), North American Business News (NAB) and Switchboard (SWBD). The transcripts of the acoustic training material (TAT) were mainly used as a cross-validation set. A model was estimated on each corpus and adapted to BN. Adaptive linear interpolation [9] is the preferred method to combine several models. This reduces the trigram perplexity by 13% from 175.0 to 152.5 but WER (unfortunately) only by about 2% relative. This reduction is of the order of the 2σ error on the WER.

2. Combining Words to Phrases

In last year's evaluation it turned out that phrases may improve the performance of the recognizer [1]. However, there no automatic algorithm for selecting phrases was presented. Techniques for this goal are known from text compression [10] and applied in [2], [11] and other works for constructing phrases on small corpora. We used an algorithm that is capable of constructing phrases with a reasonable number of passes through the corpus - even for very large corpora. It is based on count or likelihood criteria and joins several phrases in one pass. For details see [12]. We present results for WSJ and BN.

For the Wallstreet-Journal Corpus, which consists of about 40 million words, the perplexities are presented in Tab. 1. The vocabulary size is 5 K words. For 226 phrases added to the vocabulary the trigram perplexity ($M = 3$) is reduced by 7.2%. Note, that all perplexities reported are normalized to words. For the fourgram the improvement is still 3.0% and only for the fivegram there is no change. As for some applications very good performance of bigrams is important we also present results for a large number of phrases. Here, using 3831 phrases, the improvement is 29%.

However, the main effect of phrases is not the reduction in perplexity but the reduction in WER. Transcriptions of the phrases were generated by an automatic transcription tool [13] and, after adding pronunciation variants by hand, added to the lexicon. Usually, phrases have more

This work was partially funded by the German Federal Ministry for Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01 IV 701 C9. The responsibility for the contents of this study lies with the authors.

# Phrases	0	226	3831
M=1	738.0	562.6	382.0
M=2	113.0	100.0	80.3
M=3	60.8	56.4	55.8
M=4	52.6	51.0	53.6
M=5	50.4	50.3	53.2

Table 1: *Perplexities for WSJ.*

pronunciation variants than ordinary words. The system was retrained with the new lexicon. Word error rates for cross-word decoding for the WSJ task are reported in Tab. 2. It is striking that there is a reduction in WER larger than expected from the perplexity reductions and is about 10 percent relative. Also, this reduction in WER persists for bigram, trigram and fourgram decoding.

Broadcast-News (BN) is the set-up we are actually aiming at. The training corpus consists of 140 million words of transcribed broadcast-news and the test set is taken from the 1996 development data. The vocabulary size is 64 K words and 330 phrases are constructed. Perplexities are shown in Tab. 3. The improvement for bigrams is 8.4% and for trigrams 4.1%.

The most frequent phrases are given in Tab. 4. Also their position in a frequency sorted vocabulary is reported. The first ten phrases are among the 90 most frequent words. This leads to a reduction of the number of events in the training corpus by about 10 percent. However, this does not lead to problems when estimating language models as the phrases are very frequent and hence transition probabilities are well estimated. Those ten phrases are also occurring in WSJ except for "you_know" and "I_think" which seem not to be common in newspapers. Some of the phrases reported in [1] are not found like "What_did_you", because there manual selection and different criteria have been used. For other phrases from [1], we also found many related variants. Not only "going_to" was found but also: "going_to_be", "we're_going_to", "going_to_have", "is_going_to", "not_going_to" and "are_going_to".

Model	M=2	M=3	M=4
5 K words	8.2%	7.0%	6.9%
+ 226 phrases	7.7%	6.1%	6.0%

Table 2: *Word error rate on WSJ for bigram, trigram and fourgram language models showing the influence of phrase construction.*

	M=1	M=2	M=3
64 K words	1026.4	257.1	180.0
+ 330 phrases	841.2	235.4	172.7

Table 3: *Perplexities for BN.*

Joining phrases increases the effective length of the words in a selective manner. Using 226 phrases on WSJ the average length (i.e. weighted by frequency of the word or phrase) of the new words in terms of the old words is 1.13. When constructing 3831 phrases this value increases to 1.35. For BN we have 1.16. This shows the limits of the method. The average unigram context is still quite short and M-gram models are necessary to model the relation between the new basic units. However, there are also a few examples of very long phrases. For the set-up with 3831 phrases the longest one (with 1117 occurrences in the corpus!) is "in_New_York_stock_exchange_composite_trading_yesterday".

3. FMA: An Adaptation Technique Based on Unigram Distributions

3.1. Adapted Marginals as Constraints

The goal of this section is to review a method that allows adaptation of NAB to the BN domain. The method of fast marginal adaptation (FMA) has already been presented in great detail in [6] and we only want to give a brief summary and an alternative point of view.

Unigram distributions are reliably estimated. Hence, it is very desirable to use this information as a constraint when adapting a model to a different domain. Thus,

Phrase	Position
in_the	14
of_the	17
on_the	45
to_the	46
and_the	64
you_know	70
for_the	76
to_be	79
I_think	81
that_the	88

Table 4: *Most frequent phrases for BN and their position in a frequency sorted vocabulary.*

	Perplexity								WER
Condition	F0	F1	F2	F3	F4	F5	FX	F0 - FX	F0 - FX
NAB	378.1	377.6	459.8	453.7	432.4	360.9	344.0	395.2	43.4%
NAB adap.	360.8	249.7	286.8	410.4	396.4	335.6	276.0	313.4	43.0%
BN	318.9	165.9	179.6	345.8	366.8	294.9	213.2	241.9	42.0%
BN + TAT	314.0	163.7	176.5	332.8	369.3	293.0	211.9	239.0	41.8%
BN + TAT + NAB adap.	286.4	157.0	167.3	306.6	338.3	278.7	202.2	224.3	41.2%

Table 5: *Changes in perplexity and WER by adaptation of bigram models.*

we consider $P_{BN}(w)$ and $P_{NAB}(\mathbf{h}\, w)$ as given and an unknown $P_{Adap}(\mathbf{h}\, w)$ will be determined. Summing over all histories $\mathbf{h}$, this gives a constraint

$$\sum_{\mathbf{h}} P_{Adapt}(\mathbf{h}\, w) = P_{BN}(w) \qquad (1)$$

At the same time we require the Kullback–Leibler distance

$$D(P_{Adapt}||P_{NAB}) \qquad (2)$$

to be minimal.

This problem can be solved be generalized iterative scaling (GIS) [8]. Instead of attempting a converged iterative solution of the problem we just employ the first iteration step of the GIS algorithm to obtain a closed solution. However, up to now it has not really been mentioned in the literature that for every iteration step, there is a free parameter to optimize the quality of the step. Using this, the result is

$$P_{Adapt}(w|\mathbf{h}) = \frac{1}{Z(\mathbf{h})}\left(\frac{P_{BN}(w)}{P_{NAB}(w)}\right)^{\beta} P_{NAB}(w|\mathbf{h}) \qquad (3)$$

where β is the free parameter, which controls the convergence properties. The optimal numerical value is determined on a cross-validation set and $Z(\mathbf{h})$ is the normalization, which can be efficiently calculated as shown in [6] to allow for the use of this algorithm in speech recognition.

The marginal unigram distribution is estimated on BN. The background model is the bigram estimated on the 240 million words of NAB. Adaptation results on the test set of the 1996 evaluation development data are summarized in Tab. 5. Results for the adaptation itself (first two rows) and how it influences the combined model (last three rows) are given. In the table “NAB” refers to the NAB bigram and “NAB adap.” to the bigram adapted to the BN domain using FMA.

First, we observe, that on F1 and F2 (those two conditions are spontaneous speech) the improvement by adaptation is largest, as they are very remote from the NAB domain. The reduction of bigram perplexity for the NAB based model only, is 34% on F1 and 38% on F2. Averaged over all conditions the improvement is 21%. Also the WER is reduced.

Combining (by linear interpolation) the adapted model and the already existing model, which is trained on BN and TAT, gives the largest improvement on the planned speech conditions like F0, but there is also a reduction in perplexity for F1 and F2. In addition, there is again a small decrease in WER.

3.2. Complete NAB model

The previous section described the FMA technique for adaptation. Now, a complete fourgram model that is based on NAB and adapted to BN will be described (Tab. 6). Various parts are combined by linear interpolation. The fourgram models have been pruned such as to reduce the loss in information as described in [3]. To make up for the inevitable loss due to the pruning, trigrams, distance-2 and distance-3 bigrams are added to the model. The weights for the linear interpolation are given in Tab. 6. They have been optimized on TAT as a cross validation set. It is interesting to observe that the models with a true fourgram context have only a combined weight of 0.363. Thus, there seems to be still room for improvement in modeling fourgram models. For the trigram, the adapted model clearly dominates. For the distance bigrams, the non-adapted models were removed from the combined model as they had negligible weights.

4. Combination of BN, BNC, NAB, SWBD and TAT.

BNC, NAB and SWBD are used, adapted and smoothed as described in the previous section. They are combined by adaptive linear interpolation [9]. The initial interpolation weights are given in Tab. 7. As the evaluation is unpartitioned the cross-validation set on which the model is optimized is a mixture of all conditions. The weight of BN is quite large, and NAB is the only one that seems to contribute to the combined model. However, when looking at the improvement in perplexity for

NAB submodel	Weight
M=4	0.179
M=4 adapted	0.184
M=3	0.102
M=3 adapted	0.414
M=2 d=2 adapted	0.040
M=2 d=3 adapted	0.081

Table 6: *Weights for the NAB-model*

the different conditions, it is observed that SWBD improves the model for F1 and F2 and BNC improves it for F5. As we use adaptive linear interpolation, the initial parameters get changed and the proper submodel obtains a larger weight.

To be more specific, in Tab. 8 perplexities and WER are reported and compared for trigrams and fourgrams. For the trigram, perplexity is reduced by 13% by adding the other corpora. The WER is reduced by moderate 2% relative. For the fourgram, the PP reduction is 14% but the best fourgram is only 3% better than the best trigram. Here, no recognition experiments have been performed.

Corpus	BN	NAB	BNC	SWBD
Weight	0.801	0.150	0.032	0.017

Table 7: *Weights for the different corpora for the fourgram model.*

Model	PP	WER
M=3 BN + TAT	175.0	39.4%
M=3 all	152.5	38.6%
M=4 BN	171.0	-
M=4 all	147.7	-

Table 8: *Results for the combination of BN, TAT, NAB, BNC and SWBD (referred to as "all")*

5. Results on the Evaluation Data

This final section briefly summarizes key figures for the 1997 evaluation data and also summarizes the paper. Two different vocabularies have been tested. The first one is based on the 64 K most frequent words of BN. This vocabulary has an out-of vocabulary (OOV) rate of 0.49%. The second vocabulary was supplemented by 831 words from TAT and the speaker data base. This yields an OOV rate of 0.48%. Tab. 9 gives the perplexities for the language models described above on the 1997 data. For the bigram, phrases give a 9% improvement and the adaptive combination again 9%. Adding TAT to the trigram gives a small improvement but when combining all corpora perplexity is reduced by 15%. The best fourgram is insignificantly better than the best trigram. This is probably because we consistently use phrases which make the trigram actually a 3.5-gram and the fourgram a 4.6-gram. The models marked with an asterisk in the table have actually been used in the evaluation.

Model	PP
M=2 BN without phrases	236.3
M=2 BN	215.7
M=2 BN + TAT + NAB adap. (*)	195.9
M=3 BN	149.9
M=3 BN + TAT (*)	146.6
M=3 all corpora	127.4
M=4 BN	144.4
M=4 all corpora	125.8

Table 9: *Perplexities on the 1997 Evaluation Data*

6. Conclusion

The experiments with phrases for BN indicate, that they already cover a long context, Hence, the additional gain when going from a trigram to a fourgram is small. Also, adapting remote corpora to BN yields reductions in perplexity and WER.

7. Acknowledgment

We would like to thank Stefan Besling, Reinhard Kneser and Jochen Peters for many stimulating discussions.

References

1. J.L. Gauvain, G. Adda. L. Lamel, and M. Adda-Decker: "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System", *DARPA Speech Recognition Workshop*, pp. 56, 1997.
2. K. Ries, F. D. Buo, and A. Waibel: "Class Phrase Models for Language Modeling", *Proc. ICSLP*, pp. 398, 1996.
3. R. Kneser: "Statistical Language Modeling Using a Variable Context Length", *Proc. ICSLP*, pp. 494, 1996.
4. S. Della Pietra, V. Della Pietra, R.L. Mercer and S. Roukos: "Adaptive Language Modeling using Minimum Discriminant Estimation",*Proc. ICASSP*, pp. 663, 1992.
5. P. Srinivasa Rao, M.D. Monkowski, and S. Roukos: "Language Model Adaptation via Minimum Discriminant Information",*Proc. ICASSP*, pp. 161, 1995.
6. R. Kneser, J. Peters and D. Klakow: "Language Model Adaptation Using Dynamic Marginals",*Proc. EUROSPEECH*,pp. 1971, 1997.

7. P. Srinivasa Rao, Satya Dharanipragada, Salim Roukos: "MDI Adaptation of Language Models Across Corpora",*Proc. EUROSPEECH*, pp. 1979, 1997.

8. J.N. Darroch and D. Ratcliff: "Generalized Iterative Scaling for Log-Linear Models",*The Annals of Mathematical Statistics*,pp. 1470, 1972.

9. R. Kneser and V. Steinbiß: "On the Dynamic Adaptation of Stochastic Language Models",*Proc. ICASSP*, pp. 586, 1993.

10. J.A. Storer: "Data Compression", *Computer Science Press* , 1988.

11. K. Hwang: "Vocabulary Optimization based on Perplexity", *Proc. ICASSP*, pp. 1419, 1997.

12. D. Klakow: "Language-Model Optimization by Mapping of Corpora",*Proc. ICASSP*, accepted for publication, 1998.

13. S. Besling: "A Statistical System for Grapheme–to–Phoneme Conversion", *Proc. Tenth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research: Reflections on the Future of Text*, pp. 5, 1994.